# Lead Score Case Study

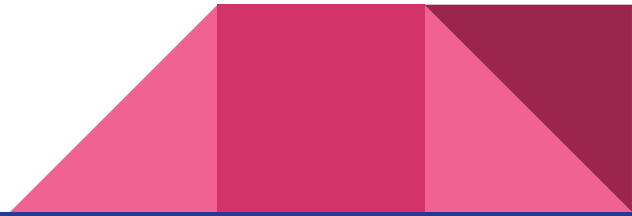
Rajani Sharma

# Problem Statement

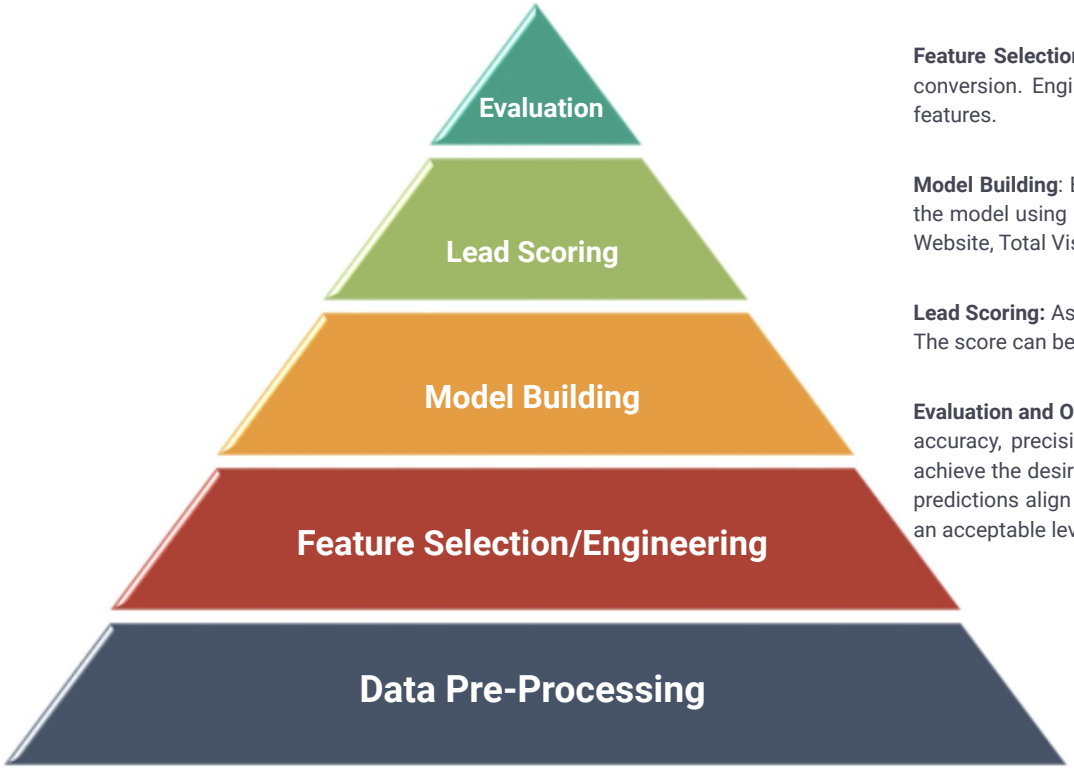
The problem is to build a lead scoring model for X Education to identify potential leads that are most likely to convert into paying customers. The company wants to improve its lead conversion rate and focus their efforts on the leads with the highest conversion potential. The goal is to assign a lead score between 0 and 100 to each lead, where a higher score indicates a higher likelihood of conversion.

# Business Objective

The primary objective is to increase the lead conversion rate from the current 30% to the target rate of around 80%. This can be achieved by accurately identifying the most promising leads, enabling the sales team to focus on nurturing and converting them. The lead scoring model will help prioritize leads and optimize the sales efforts, ultimately leading to better utilization of resources and improved business performance.



# Implementation




**Data Preprocessing**: Clean and preprocess the provided leads dataset. Handle missing values, handle 'Select' levels in categorical variables as null values, and perform necessary data transformations.

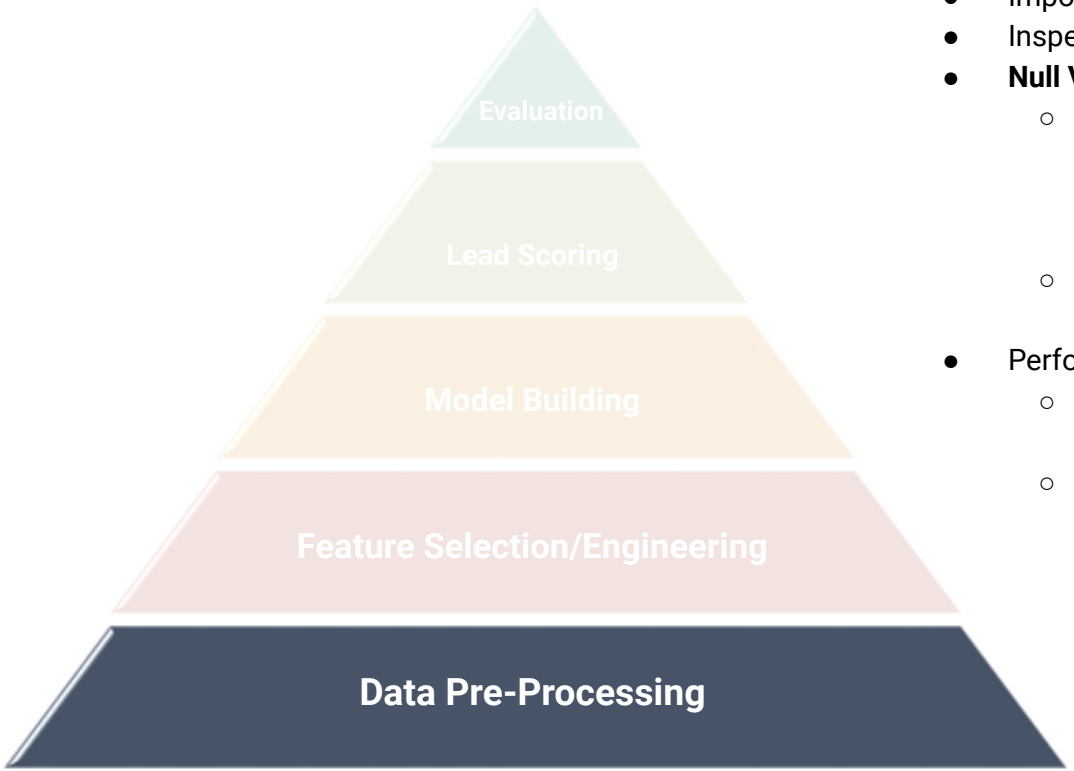
**Feature Selection**/Engineering: Identify relevant features that contribute significantly to lead conversion. Engineer new features if necessary, and remove irrelevant or highly correlated features.

**Model Building**: Build a logistic regression model to predict lead conversion probability. Train the model using historical data, including features such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc.

**Lead Scoring:** Assign a lead score to each lead based on the predicted conversion probability. The score can be scaled to a range between 0 and 100 for better interpretation.

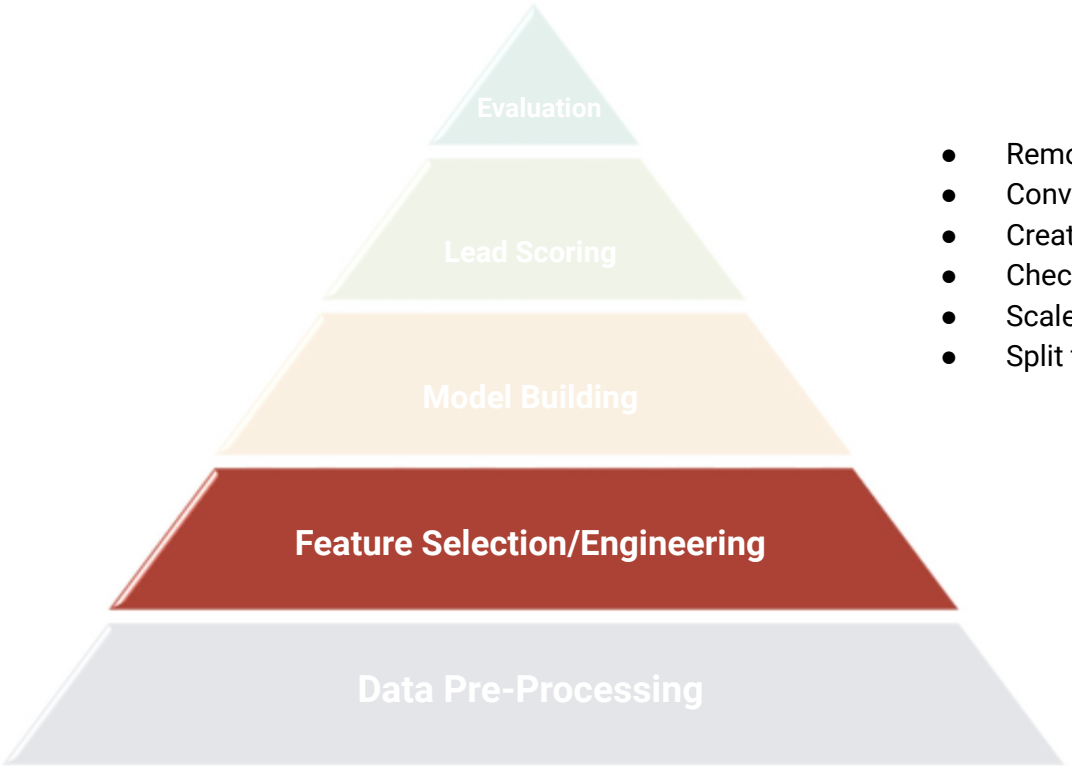
**Evaluation and Optimization:** Evaluate the model's performance using appropriate metrics like accuracy, precision, recall, and ROC curve. Optimize the model's threshold if necessary to achieve the desired trade-off between precision and recall. This step ensures that the model's predictions align with the business objectives, i.e., maximizing conversions while maintaining an acceptable level of false positives.

# Data Pre-Processing



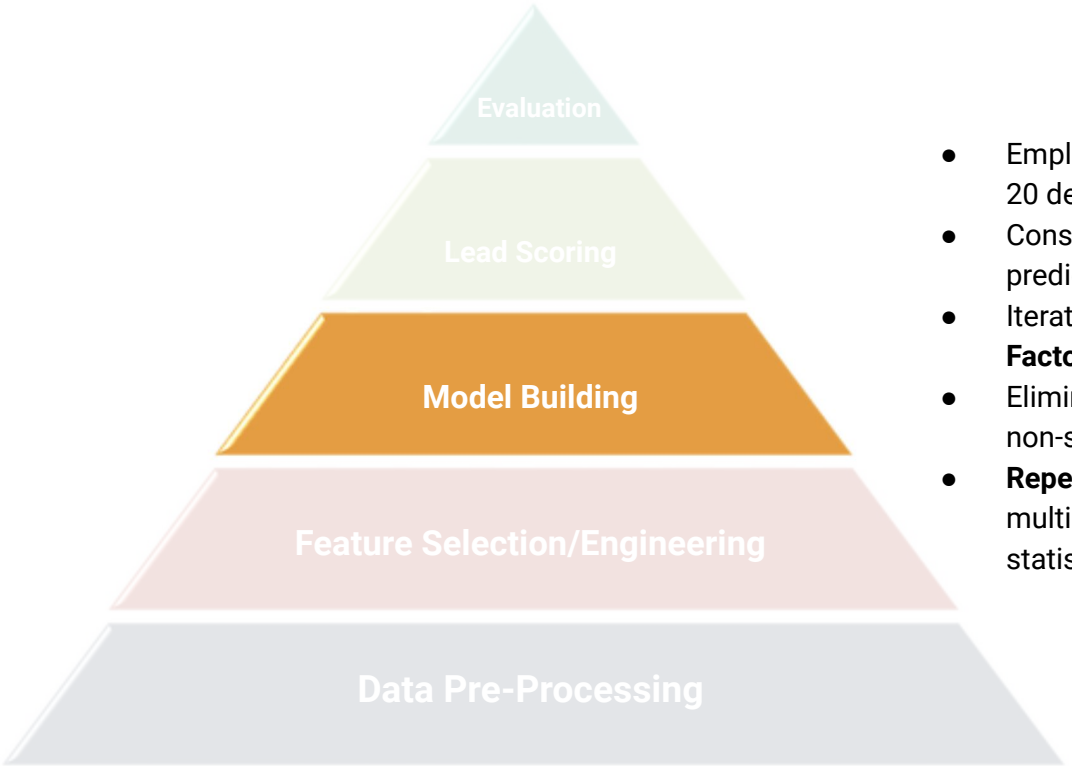

- Import the ‘leads.csv’ data
- Inspect the dataset
- **Null Values:**
  - Columns:
    - Drop columns where more than 3K(~30%) rows have missing value
    - Drop columns with highly skewed data
  - Rows:
    - Remove rows with any null values(~2%)
- Perform necessary data transformations on **outliers**
  - Categorical variable: Replace ‘Select’ and nulls with ‘Others’
  - Numerical variable: Replace outliers with 95%ile value

# Feature Selection/Engineering



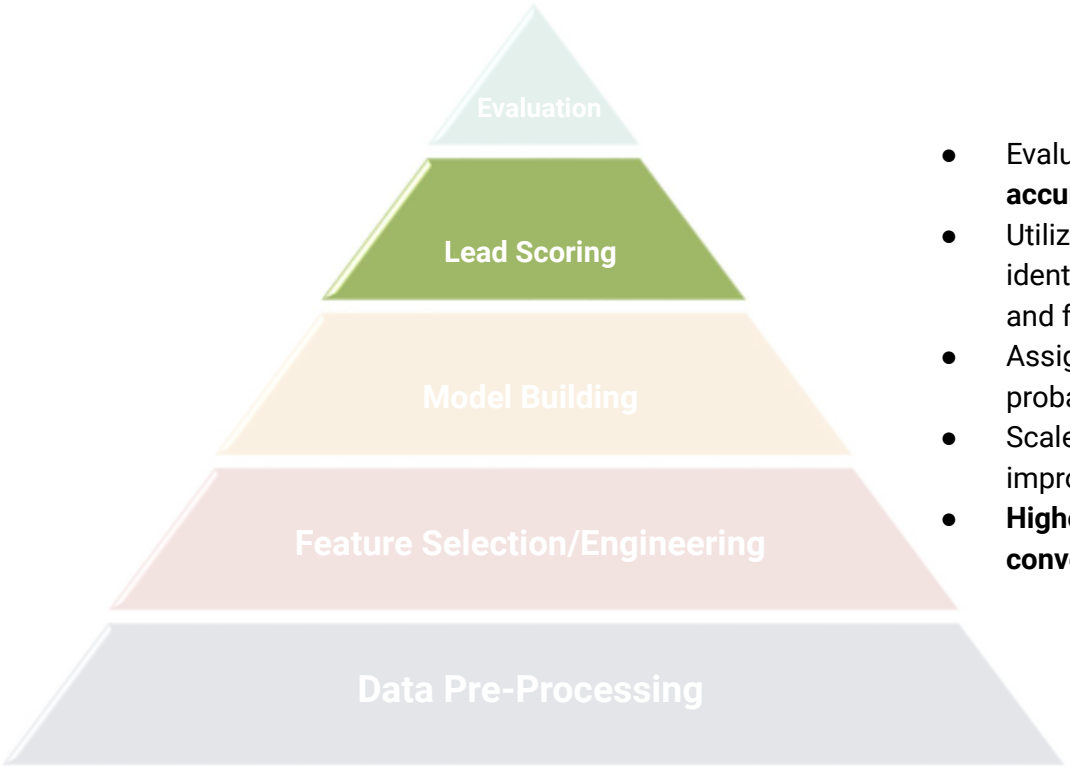

- Remove irrelevant features i.e., IDs
- Converting some binary variables (Yes/No) to 1/0
- Creating **Dummy variables** for the categorical features
- Check correlations
- Scale the numerical columns using **MinMaxScaler**
- Split the data in 80-20 ratio for **train-test** sets

# Model Building



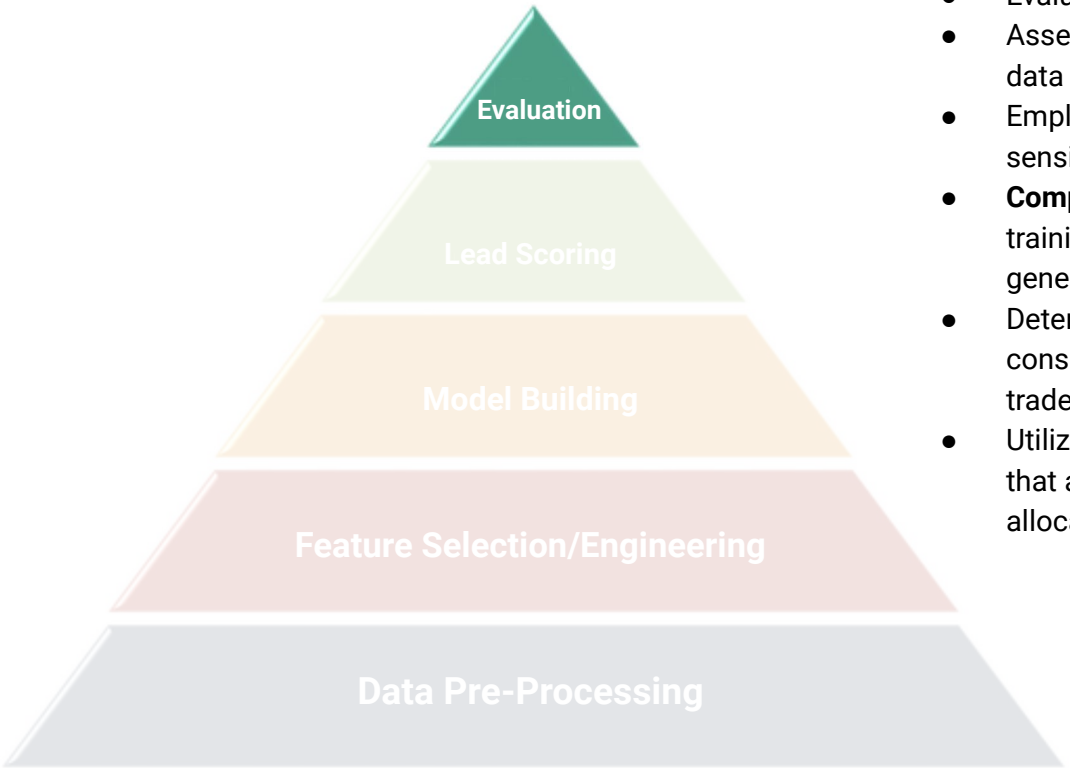

- Employ **Recursive Feature Elimination (RFE)** to select the top 20 dependent variables.
- Construct a **logistic regression model** for lead conversion prediction using the training data.
- Iteratively eliminate columns with high **Variance Inflation Factor (VIF)** to mitigate multicollinearity.
- Eliminate columns with high **p-values** to address non-significant predictors.
- **Repeat** the model-building process iteratively until multicollinearity is minimized and all remaining columns are statistically significant.

# Lead Scoring



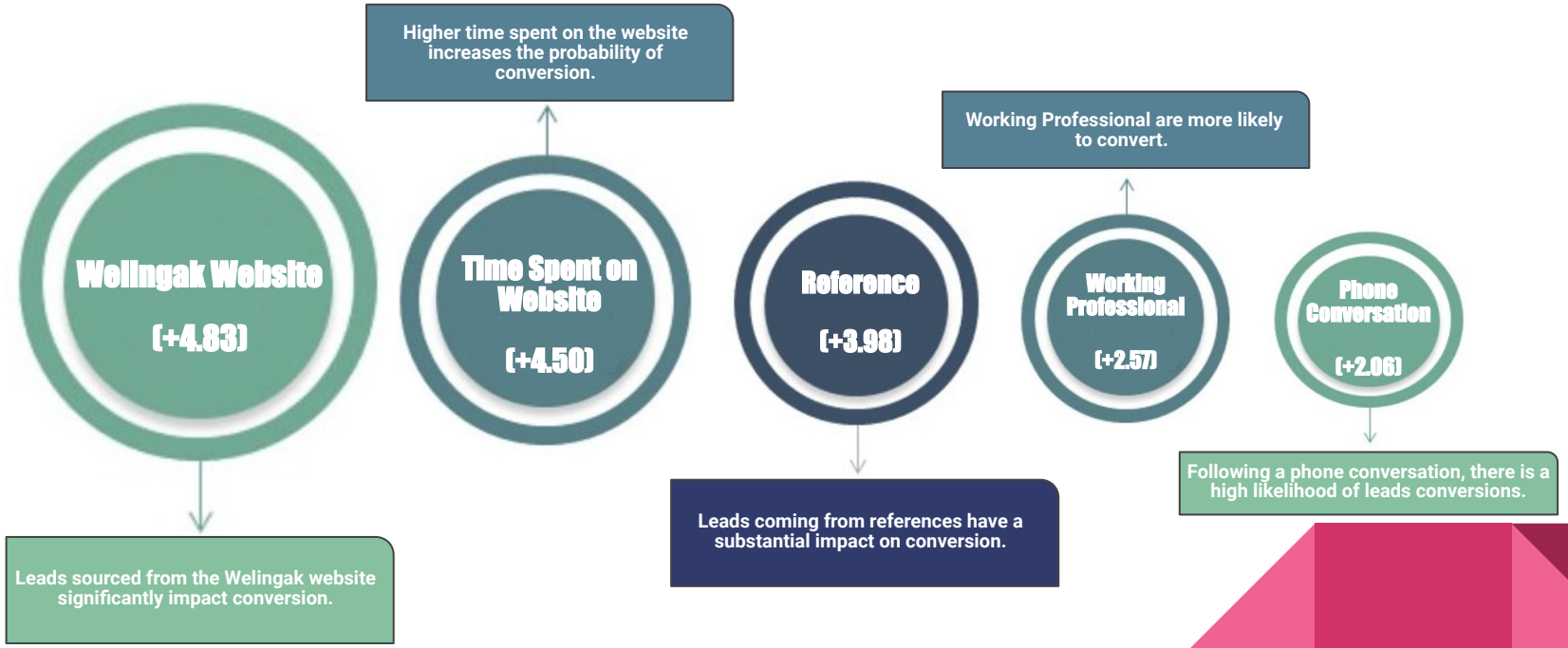

- Evaluate model performance using various metrics including **accuracy, specificity, and sensitivity.**
- Utilize **Receiver Operating Characteristic (ROC)** curve to identify the optimal cutoff point for balancing true positives and false positives.
- Assign lead scores based on the model's predicted conversion probabilities.
- Scale the lead scores to a range between **0 and 100** for improved interpretation.
- **Higher lead scores indicate a stronger likelihood of lead conversion**, aiding in prioritizing leads effectively.

# Evaluation on Test Data

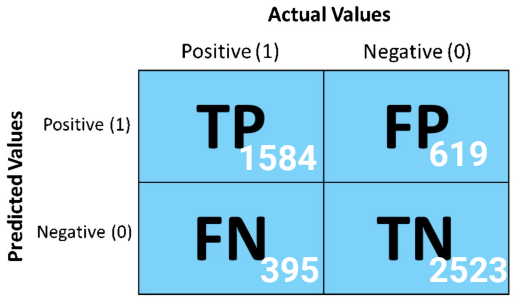



- Evaluate the trained model's performance on test data.
- Assess the performance of the trained model using the test data set.
- Employ a range of metrics including accuracy, specificity, and sensitivity to comprehensively gauge model effectiveness.
- **Compare the performance metrics** obtained from both the training and test datasets to identify potential overfitting or generalization issues.
- Determine the **optimal threshold** for making predictions, considering business objectives and precision-recall trade-offs.
- Utilize the model's predictions to identify and prioritize leads that are more likely to convert, aiding in efficient resource allocation for lead follow-up.

# Critical Indicators

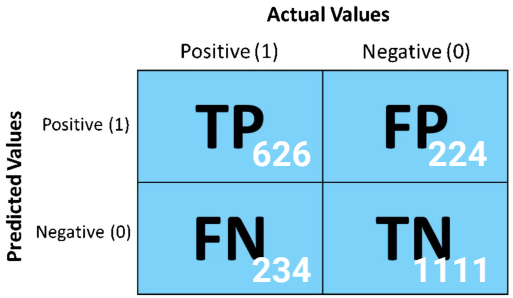


# Confusion Matrix




**Train Set**

- Accuracy: 0.80
- Sensitivity: 0.80
- Specificity: 0.80
- False Positive Rate: 0.20
- Positive Predictive Value: 0.72
- Negative Predictive Value: 0.86

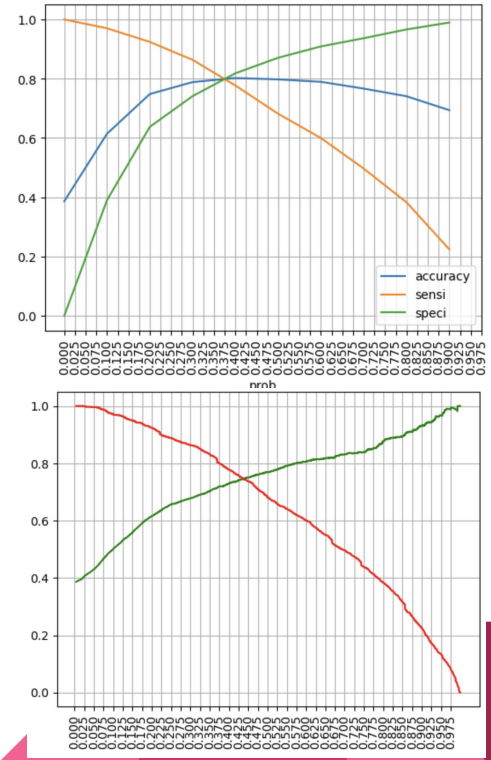



**Test Set**

- Accuracy: 0.79
- Sensitivity: 0.73
- Specificity: 0.83
- False Positive Rate: 0.17
- Positive Predictive Value: 0.74
- Negative Predictive Value: 0.83

# Threshold Optimization for Effective Lead Scoring

- **ROC Curve Analysis:**
  - Balancing true positives while maintaining overall accuracy and specificity.
  - Graphical representation guided threshold choice.
  - Identified optimal cutoff at 0.375.
  - Decision based on sensitivity, accuracy, and specificity.
- **Precision-Recall Trade-off:**
  - Optimization to balance precision and recall.
  - Precision-recall curve indicates a cut-off at 0.43.
  - Achieved balance at a threshold of 0.42 for accurate positive predictions while capturing actual positives effectively.



# Driving Conversion Success Through Informed Strategies

- **Call Leads with 85+ Score:**
  - Leads with high scores (85+) indicate strong conversion potential.
  - Prioritize these leads for immediate and focused follow-up.
  - Allocate resources efficiently by concentrating efforts on these promising prospects.
- **Aggressive Follow-up with Lower Cut-off:**
  - This approach results in more leads being classified as potential conversions.
  - Enable a proactive outreach strategy, ensuring that potentially valuable leads are not missed.
- **Enhanced Precision with Higher Cut-off:**
  - Focus on precision, ensuring that leads contacted have a high probability of converting.
  - Reduces false positive predictions, leading to more productive interactions.
- **Customize Outreach Strategies:**
  - Tailor communication strategies based on lead scores.
  - High-scoring leads can receive personalized and persuasive messages.
  - Low-scoring leads can be nurtured with informative content to improve engagement.
- **Collaboration Between Sales and Marketing:**
  - Foster collaboration between sales and marketing teams.
  - Share insights from lead scoring to align strategies.
  - Enhance lead nurturing and conversion efforts through data-driven decisions.

# Conclusion

In conclusion, our predictive model demonstrates promising performance in predicting lead conversions. With an accuracy of approximately **80% on train dataset and 79% on the test dataset**, the model showcases a notable ability to make accurate overall predictions. This accuracy is a testament to the model's ability to correctly classify both positive and negative cases.

Furthermore, key contributing dependent variables, such as **"Lead Source: Welingak Website", "Total Time Spent on Website," "Lead Source: Reference," "Current Occupation: Working Professional," and "Last Activity: Had a Phone Conversation,"** have been identified as influential factors in predicting lead conversions.

Leveraging these variables has enabled our model to effectively prioritize leads for follow-up, optimizing conversion rates and resource allocation in alignment with business objectives.